

Characterization of behavioral patterns exploiting description of geographical areas

Zolzaya Dashdorj¹ and Stanislav Sobolevsky²

¹University of Trento (Via Sommarive, 9 Povo, TN, Italy), and SKIL
LAB - Telecom Italia, and DKM - Fondazione Bruno Kessler, and
SICT - Mongolian University of Science and Technology (Bayanzurkh
district 22th khoroo, UB, Mongolia)
dashdorj@disi.unitn.it

²Massachusetts Institute of Technology (MIT 77 Massachusetts Avenue
Cambridge, MA, USA) and
New York University (1 MetroTech Center, Brooklyn, NY)
stanly@mit.edu

Abstract

The enormous amount of recently available mobile phone data is providing unprecedented direct measurements of human behavior. Early recognition and prediction of behavioral patterns are of great importance in many societal applications like urban planning, transportation optimization, and health-care. Understanding the relationships between human behaviors and location's context is an emerging interest for understanding human-environmental dynamics. Growing availability of Web 2.0, i.e. the increasing amount of websites with mainly user created content and social platforms opens up an opportunity to study such location's contexts. This paper investigates relationships existing between human behavior and location context, by analyzing log mobile phone data records. First an advanced approach to categorize areas in a city based on the presence and distribution of categories of human activity (e.g., eating, working, and shopping) found across the areas, is proposed. The proposed classification is then evaluated through its comparison with the patterns of temporal variation of mobile phone activity and applying machine learning techniques to predict a timeline type of communication activity in a given location based on the knowledge of the obtained category vs. land-use type of the locations areas. The proposed classification turns out to

be more consistent with the temporal variation of human communication activity, being a better predictor for those compared to the official land use classification.

Index terms— land-use, cell phone data records, big data, human activity recognition, human behavior, knowledge management, geo-spatial data, clustering algorithms, supervised learning algorithms

1 Introduction

Recent extensive penetration of digital technologies into everyday life have enabled creation and collection of vast amounts of data related to different types of human activity. When available for research purposes this creates an unprecedented opportunity for understanding human society directly from its digital traces. There is an impressive amount of papers leveraging such data for studying human behavior, including mobile phone records [5, 16, 29, 30, 31], vehicle GPS traces [22, 37], social media posts [20, 21, 25] and bank card transactions [38, 39]. With the growing mobile phone data records, environment modeling can be designed and simulated for understanding human dynamics and correlations between human behaviors and environments. Environment modeling is important for a number of applications such as navigation systems, emergency responses, and urban planning. Researchers noticed that type of the area defined through official land-use is strongly related with the timeline of human activity [13, 24, 28, 33, 48]. But those sources of literature do not provide extensive analyses on categorical profile of the geographical areas. This limits the understanding of the dependency of human behaviors from geographical areas. Our analysis confirms this relation, however we show that land-use by itself might be not enough, while categorical profile of the area defined based on OSM provides a better prediction for the activity timeline. For example, even within the same land-use category, timelines of activity still vary depending on the categorical profile. In this paper, different from these works, we start from clustering the entire city based on area profiles, that are a set of human activities associated with a geographical location, showing that those activities have different area types in terms of the timelines of mobile phone communication activity. Further we show that even the areas of the same land-use, which is formally defined by land-use management organizations, might have different clusters based on points of interest (POIs). But those clustered areas are still different in terms of the timelines. This will contribute to other works showing that not only the land-use matters for human activity. This paper uses mobile phone data records to determine the relationship between human behaviors and geographic area context [9]. We present a series of experimental results by comparing the clustering algorithms aiming at answering the following questions: 1). To what extent can geographical types explain

human behaviors in a city, 2). What is the relationship between human behaviors and geographical area profiles? We demonstrate our approach to predict area profiles based on the timelines of mobile phone communication activities or vice versa: to predict the timelines from area profiles. We validate our approach using a real dataset of mobile phone and geographic data of Milan, Italy. Our area clustering techniques improve the overall accuracy of the baseline to 64.89%. Our result shows that land-uses in city planning are not necessarily well defined that an area type is better defined with one type of human activity. But growing and development of city structures enable various types of activities that are present in one geographical area. So this type of analysis and its application is important for determining robust land-uses for city planning. Also the hidden patterns and unknown correlations can be observed comparing the mobile phone timelines in relevant areas. The result of this work is potentially useful to improve the classifications of human behaviors for better understanding of human dynamics in real-life social phenomena and to provide a decision support for stakeholders in areas, such as urban city, transport planning, tourism and events analysis, emergency response, health improvement, community understanding, and economic indicators. The paper is structured as follows Section 3 introduces the data sources we use in this research and the data-processing performed. The methodology is described in Section 4. We present and discuss the experimental results in Section 5. Finally, we summarize the discussions in Section 6.

2 Related Works

Human behavior is influenced by many contextual factors and their change, for instance, snow fall, hurricane, and festival concerts. There are number of research activities that shed new light on the influence of such contextual factors on social relationships and how mobile phone data can be used to investigate the influence of context factors on social dynamics. Researchers [2, 4, 15, 28] use an additional information about context factors like social events, geographical location, weather condition, etc in order to study the relationship between human behaviors and such context factors. This is always as successful as the quality of the context factors. The combination of some meteorological variables, such as air temperature, solar radiation, relative humidity, can effect people’s comfort conditions in outdoor urban spaces [43], poor or extreme weather conditions influence peoples physical activity [45]. Q. Wang et al. [49] exhibited high resilience, human mobility data obtained in steady states can possibly predict the perturbation state. The results demonstrate that human movement trajectories experienced significant perturbations during hurricanes during/after the Hurricane Sandy in 2012. Sagl et al. [35] introduced an approach to provide additional insights in some

interactions between people and weather. Weather can be seen as a higher-level phenomenon, a conglomerate that comprises several meteorological variables including air temperature, rainfall, air pressure, relative humidity, solar radiation, wind direction and speed, etc. The approach has been significantly extended to a more advanced context-aware analysis in [36]. Phithakkitnukoon et al. [28] used POIs to enrich geographical areas. The areas are connected to a main activity (one of the four types of activities investigated) considering the category of POIs located within it. To determine groups, that have similar activity patterns, each mobile user's trajectory is labeled with human activities using Bayes Theorem in each time-slot of a day for extracting daily activity patterns of the users. The study shows that daily activity patterns are strongly correlated to a certain type of geographic area that shares a common characteristic context. Similar to this research idea, social networks [48] have been taken into account to discover activity patterns of individuals. Noulas et al. [24] proposed an approach for modelling and characterization of geographic areas based on a number of user check-ins and a set of eight type of general (human) activity categories in Foursquare. A Cosine similarity metric is used to measure the similarity of geographical areas. A Spectral Clustering algorithm together with K-Means clustering is applied to identify an area type. The area profiles enables us to understand groups of individuals who have similar activity patterns. Soto and Frias-Martinez et al. [42] studied mobile phone data records to characterize geographical areas with well defined human activities, by using the Fuzzy C-Means clustering algorithm. The result indicated that five different land-uses can be identified and their representation was validated with their geographical localization by the domain experts. Frias-Martinez et al. [13] also studied geolocated tweets to characterize urban landscapes using a complimentary source of land-use and landmark information. The authors focused on determining the land-uses in a specific urban area based on tweeting patterns, and identification of POIs in areas with high tweeting activity. Differently, Yuang et al. [50] proposed to classify urban areas based on their mobility patterns by measuring the similarity between time-series using the Dynamic Time Warping (DTW) algorithm. Some areas focus on understanding urban dynamics including dense area detection and their evolution over time [23, 46]. Moreover, [14, 32, 41] analyzed mobile phone data to characterize urban systems. More spatial clustering approaches (Han & Kamber [19]) could group similar spatial objects into classes, such as k-means, k-medoids, and Self Organizing Map. They have been also used for performing effective and efficient clustering. In this research, we use spectral clustering with eigengap heuristic followed by k-means clustering. Calabrese et al. [33] and also [18, 27] used eigengap heuristic for clustering urban land-uses. In many works [3, 26, 32, 34, 40, 44] the authors analyzed mobile phone data activity timelines to interpret land-use type. T. Pei et al. [26] analyzed the correlation between

urban land-use information and mobile phone data. The author constructed a vector of aggregated mobile phone data to characterize land-use types composed of two aspects: the normalized hourly call volume and the total call volume. A semi-supervised fuzzy c-means clustering approach is then applied to infer the land-use types. The method is validated using mobile phone data collected in Singapore. land-use is determined with a detection rate of 58.03%. An analysis of the land-use classification results shows that the detection rate decreases as the heterogeneity of land-use increases, and increases as the density of cell phone towers increases. F. Girardin et al. [17] analyzed aggregate mobile phone data records in New York City to explore the capacity to quantify the evolution of the attractiveness of urban space and the impact of a public event on the distribution of visitors and on the evolution of the attractiveness of the points of interest in proximity.

3 Collecting and Pre-Processing the data

We use two types of datasources for this experiment; 1) POIs from available geographical maps, Openstreetmap 2) Mobile phone network data (sms, internet, call, etc) generated by the largest operator company in Italy. The mobile phone traffic data is provided in a spatial grid, the rectangular grid of 100 x 100 square of dimension 235 m x 235 m . We use the grid as our default configuration for collecting human activity distribution and mobile network traffic activity distribution.

3.1 Point of Interests extracted from Openstreetmap

In [6, 7, 8, 11], one of the key elements in the contextual description of geographical regions is the point of interest (POI) (e.g. restaurants, ATMs, and bus stops) that populates an area. A POI is a good proxy for predicting the content of human activities in each area that was well evaluated in [10]. Employing a model proposed in [10], a set of human activities likely to be performed in a given geographical area, can be identified in terms of POI distribution. This allows us to create area profiles of geographical locations in order to provide semantic (high level) descriptions to mobile phone data records in Milan. For example, a person looking for food if the phone call is located close to a restaurant. We exploit the given spatial grid to enrich the locations with POIs from open and free geographic information, Openstreetmap (OSM)¹. We collected in total 552,133 POIs that refined into 158,797 activity relevant POIs. To have a sufficient number and diversity of POIs in each location, we consider the nearby areas for estimating the likelihood of human activities. The nearby areas are the intersected

¹<http://www.openstreetmap.org>

locations within the aggregation radius of the centroid point at each location. The aggregation radius is configured differently in each location, which satisfies the need for the total number of POIs in such intersected locations to be above the threshold h , see Figure 1(b) and 1(a) where each location at least $h=50$ POIs in the intersected locations. Across locations, the min, median, and max number of POIs are 50, 53, and 202.

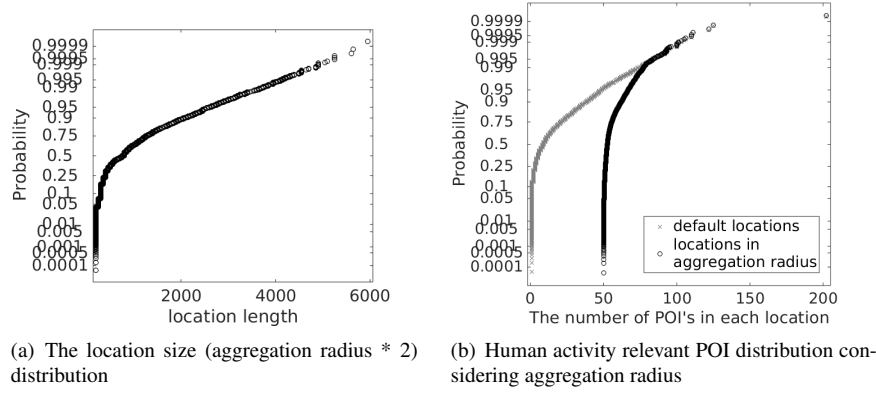


Figure 1: The distributions of POIs and human activities across locations

In order to build area profiles of each location, a $n \times m$ dimensional matrix $A_{n,m}$ is defined for each location $n \in \{1, \dots, 10000\}$. Each element $A_{n,m}$ contains the weight of activity categories m in location n where the $m \in \{\text{eating, educational, entertainment, health, outdoor, residential, shopping, sporting, traveling, working}\}$, with the total number of 10 measurements of human activities per each location. The weight of each category of activities are estimated by the HRBModel which allows us to generate a certain weight for human activities that is proportional to the weight of relevant POIs located in each location. The weight of POIs in a given location, is estimated by the following equation of $tf-idf(f, l) = \frac{N(f, l)}{\arg\max_w \{N(w, l) : w \in l\}} * \log \frac{|L|}{|\{l \in L : f \in l\}|}$, where f is a given POI; $f \in F$, $F = \{\text{building, hospital, supermarket, ...}\}$ and l is a given location; $l \in L$, $L = \{\text{location1, location2, location3, ...}\}$, $N(f, l)$ is the occurrence of POI f and its appearance in location l and $\arg\max_w \{N(w, l) : w \in l\}$ is the maximum occurrence of all the POIs in location l , $|L|$ is the number of all locations, $|\{l \in L : f \in l\}|$ is the number of locations where POI f appears.

The activity distribution in Milan area is shown in Figure 2. The sporting, working, eating and transportation types of activities are mainly performed in the city.

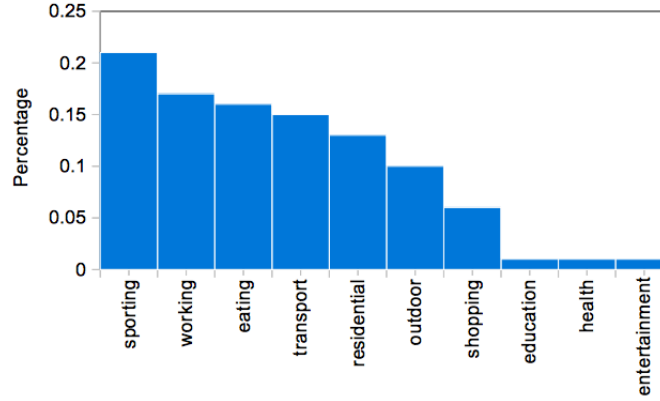


Figure 2: The activity distribution in Milan

3.2 Mobile phone network traffic

In this work, we used a dataset from “BigDataChallenge”² organized by Telecom Italia. The dataset is the result of a computation over the Call Detail Records (CDRs) generated by the Telecom Italia cellular network within Milan. The dataset covers 1 month with 180 million mobile network events in November, 2014 as November is a normal month without any particular events organized in Milan. The CDRs log the user activity for billing purposes and network management. There are many types of CDRs, for the generation of this dataset we considered those related to the following activities: square id (the id of the square that is part of the Milan GRID which contains spatially aggregated urban areas), time interval (an aggregate time), received SMS (a CDR is generated each time a user receives an SMS), sent SMS (a CDR is generated each time a user sends an SMS), incoming Calls (a CDR is generated each time a user receives a call), outgoing Calls (a CDR is generated each time a user issues a call), internet (a CDR is generated each time, a user starts an internet connection, or a user ends an internet connection).

By aggregating the aforementioned records, this dataset was created that provides mobile phone communication activities across locations. The call, sms and internet connection activity logs are collected in each square of the spatial grid for Milan urban area. The activity measurements are obtained by temporally aggregating CDRs in time-slots of ten minutes. But the temporal variations make the comparison of human behaviors more difficult. The standard approach to account for temporal variations in human behavior is to divide time into coarse grained time-slots. In Farrahi and Gatica-Perez et al. [12], the following eight coarse-grained time-slots are introduced: [00-7:00 am.,

²<http://www.telecomitalia.com/tit/it/bigdatachallenge.html>

7:00-9:00 am., 9:00-11:00 am., 11:00 am.-2:00 pm., 2:00-5:00 pm., 5:00-7:00 pm., 7:00-9:00 pm., and 9:00 pm.-00 am.]. Here, we aggregate the mobile phone network data in such coarse-grained time-slots to extract the pattern of 1 month network traffic volume in each location. For each location, we then aggregated the total number of call (outgoing and incoming call without considering a country code), and sms (incoming and outgoing), internet activity for each of those eight time-slots. Such time-slot based timelines can give us actual patterns of mobile network traffic activity.

Then the dataset reduced to 2.4 million CDR each of which consists of the followings: square id, day of month, time-slot, and total number of mobile network traffic activity. We build a $n \times p \times d$ dimensional matrix $T_{n,p,d}$ to collect a mobile phone traffic activity timeline, where n is the number of locations in [1,10000], p is the time-slot divisions of the day [1,8] and d is the day in [1,31]. To identify timeline patterns among those locations, we performed a normalization for the timelines based on z-score which transforms the timeline into the output vector with mean $\mu=0$ while standard deviation σ is negative if it is below the mean or positive if it is above the mean. The normalized timelines by day are visualized in Figure 3 which show a stable communication activity within the month. For this transformation, we used $T'_{i,j,k} = \frac{T_{i,j,k} - \mu_i}{\sigma_i}$, $i \in n, j \in p, k \in d$, where μ_i is the average value of the mobile phone activity traffic in location i , σ_i is the standard deviation of the mobile phone activity traffic in location i .

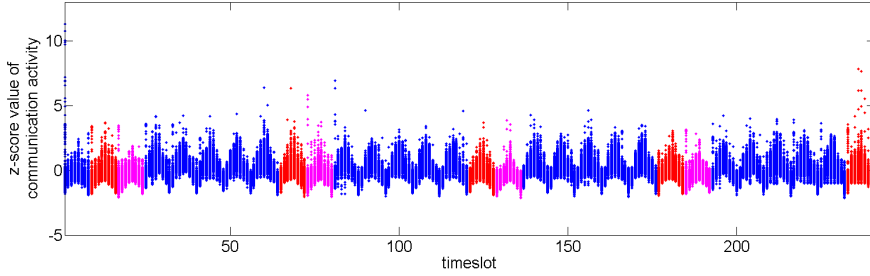


Figure 3: The timelines for each time-slot of day normalized by z-score in each location (weekday-blue, saturday-red, sunday-pink)

4 The Approach

We present our methodology for identifying the relation between geographical locations and human behaviors. Our methodology is divided into two phases: 1) clustering approaches for inferring categorical area types in terms of geographical area profiles 2)

classification approaches for validating the observed area types by mobile phone data records. Clustering techniques are mostly unsupervised methods that can be used to organize data into groups based on similarities among the individual data items. We use the spectral clustering algorithm which makes use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. The similarity matrix is provided as an input and consists of a quantitative assessment of the relative similarity of each pair of points in the dataset.

We define a vector space model that contains a set of vectors corresponding to areas. Relevance between areas is a similarity comparison of the deviation of angles between each area vector. The similarity between the areas is calculated by the cosine similarity metric by estimating the deviation of angles among area vectors. For example, the similarity between area l_1 and l_2 would be $\cos \theta_{l_1, l_2} = \frac{l_1 \cdot l_2}{\|l_1\| \|l_2\|}$ where l_i denotes the area or the features associated to the areas. We denote each area l_i with a set of corresponding features associated with a weight measure j . Having the estimation of similarity between the areas, we can now create a similarity graph described as the weight matrix W generated by the cosine similarity metrics and the diagonal degree matrix D is utilized by the spectral clustering algorithm which is the one of the most popular modern clustering methods and performs better than traditional clustering algorithms. We create the adjacency matrix A of the similarity graph and graph Laplacian LA , $LA = D - A$ (given by normalized graph Laplacian $LA_n = D^{-1/2} L A D^{-1/2}$). Based on eigengap heuristic [47], we identify the number of clusters by k-nearest neighbor to observe in our dataset as $k = \operatorname{argmax}_i (\lambda_{i+1} - \lambda_i)$ where $\lambda_i \in \{l_1, l_2, l_3, \dots, l_n\}$ denotes the eigenvalues of l_n in the ascending order. Finally, we easily detect the effective clusters (area profiles) $S_1, S_2, S_3, \dots, S_k$ from the first k eigenvectors identified by the k-means algorithms.

We investigate the relation between geographical locations and human behaviors based on categorical area types. To do that, we use supervised learning algorithms to predict area profile of a given area if we train a classification model with training data, which are the timelines labeled with area types. In supervised learning, each observation has a corresponding response or label. Classification models learn to predict a discrete class given new predictor data. We use several of classifiers for learning and prediction. We prepare a test set for testing classification models by k-fold cross validation method.

5 Experiments and Results

In this section, we demonstrate the identification of the relationships between locations and human behaviors in terms of two types of features in each location: 1) location

contexts: categories of human activity estimated through types of available POI 2) mobile communication activity timeline: mobile communication activity in time-series of coarse grained time-slots. In other words, the extent to which human behaviors depend on geographical area types. To identify and quantify these dependencies, we perform two types of validations: 1) observed area type we defined vs human behavior 2) land-use type defined formally vs human behavior by estimating the correlations and prediction algorithms.

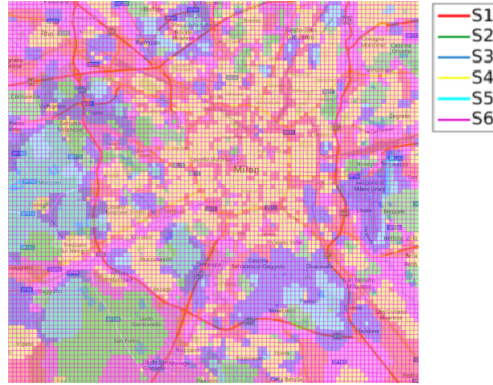


Figure 4: Observed area types of the geographical area of Milan based on the area profiles, $k=6$, where $S1$ is red, $S2$ is lime, $S3$ is blue, $S4$ is yellow, $S5$ is cyan/aqua, and $S6$ is magenta/fuchsia

5.1 Observed area type vs human behavior

We first check the two datasets can be clustered or randomly distributed using Hopkins statistic, $H = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i}$. The distance between element p_i and its nearest neighbor in dataset D is $x_i = \min_{v \in D} \{dist(p_i, v)\}$ and the distance between element q_i and its nearest neighbor in $D - q_i$ is $y_i = \min_{v \in D, v \neq q_i} \{dist(q_i, v)\}$. The Hopkins statistic for the location context dataset is 0.02 and the mobile communication timeline is 0.04 that indicates that the datasets are highly clustered and regularly distributed. So we then analyze the correlations of location context and mobile phone communication timeline in order to understand if humans are attracted to location contexts through the area types (i.e., shopping, working, and studying). To validate such relationship, we start with the geographical area clustering based on the location context by semi-supervised learning algorithms. We perform spectral clustering on the locations based on their similarity of human activity distribution $A_{n,m}$. Each location of the grid has a distribution of activity categories with relative frequency of their appearance. The

spectral clustering with k -nearest neighbor ($k = 10$ based on cosine similarity metrics) approach allows us to classify geographical areas L based on such multi-dimensional features, $A_{n,m}$. We then observed significantly different six types of areas, that are geo-located in Figure 4. The average values of the activity categories for those area types are presented in Figure 5.

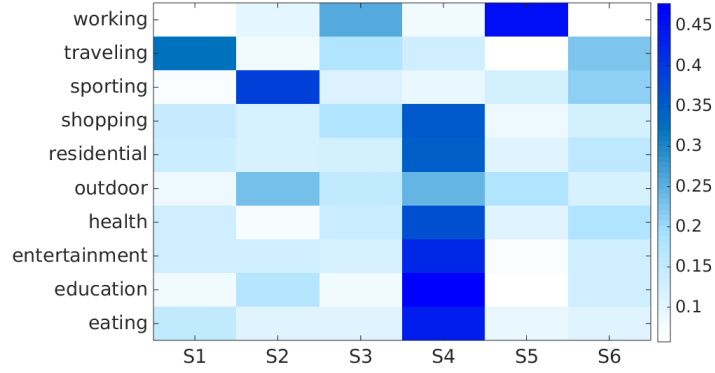


Figure 5: The average values of the activity categories in categorical area types observed

The figure shows that categorical area type $S4$ contains high percentage values for residential, and eating activities. The center of the city including a residential zone were clustered into one area type. The area type $S3$ contains high percentage value on working activity. This classification can be refined if we increase the number of area types observations. For each area type, we are now able to extract and observe timelines $T_{n,p,d}$ from mobile phone data records in order to determine the correlation between the timelines and the area profiles for those area types.

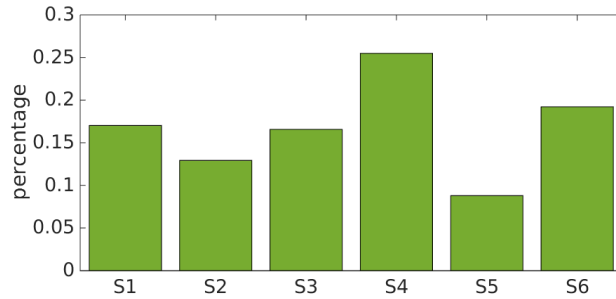


Figure 6: The density distribution of area types observed in Milan

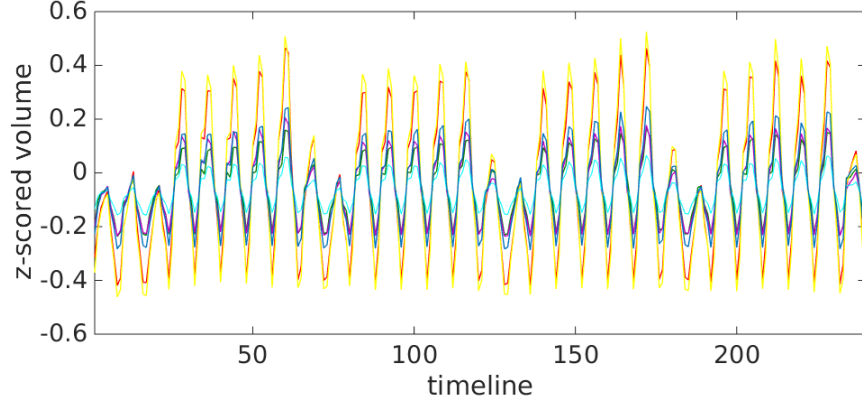


Figure 7: The average timeline of mobile phone data by area types ($k=6$), where $S1$ is red, $S2$ is lime, $S3$ is blue, $S4$ is yellow, $S5$ is cyan/aqua, and $S6$ is magenta/fuchsia

The density of the clusters are almost uniform distributed except cluster $S4$ and $S5$, see Figure 6. This unbalanced datasets for clusters could contribute to an acceptable global accuracy, but also to a (hidden) poor prediction for instances in minority classes. In this context, alternative metrics, such as per class accuracy will be considered. We estimate the accuracy per class using the two techniques (canonical correlation coefficients vs learning techniques). Figure 7 shows the actual volume of the mobile network traffic activities by the area types.

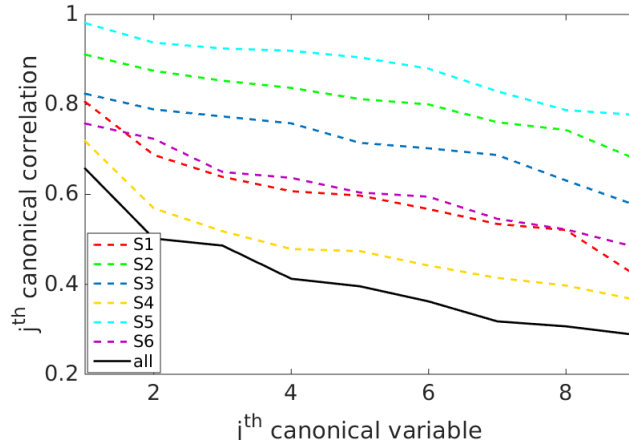


Figure 8: Canonical correlation between the two feature matrices for locations

We illustrated the correlation between the area profiles $A_{n,m}$ and timelines $T_{n,p,d}$

based on the canonical correlation [1] (see Figure 8). The canonical correlation investigates the relationships between two sets of vectors by maximizing the correlation in linear combination. In other words, canonical correlation finds the optimal coordinate system for correlation analysis and the eigenvectors define the coordinate system. While the overall maximum correlation coefficient ($j=1$) is 65% between the two vectors, the correlation coefficient by area types is high between 72% and 98%. For example, the correlation in area type *S5* is stronger than other area types, in which working type of activities are more distributed. The maximum correlation in *S2* containing high percentage of sporting activity is 82.38%.

We also compared the distance between the two vectors (mean) of area types to investigate the similarity of the relevant area profiles can have the similar human behaviors. We observed linear correlation with a coefficient of $r = 0.61$. This result shows that as the distance between the area profiles is increased, the timeline difference increases, and human behaviors are strongly correlated to geographical area profiles.

In second, we profile the communication timelines with the cluster labels observed in each location that will be used to estimate the correlation by supervised learning algorithms. The prediction accuracy of timeline types in a given location could be an evaluation of the dataset. To that end, we train several predictive models (i.e., Bayesian algorithms, Decision Trees, Probabilistic Discriminative models and Kernel machines.) to measure the prediction accuracy by k-fold cross validation method ($k=10$), which is used to estimate how accurately a predictive model will perform. We need to prepare training and test data. The training data are the timelines labeled by area types through the location. This allows us to determine if timelines are clustered as geographical area profiles. The experimental results on our data are shown in Table 1. This classification of the predictive models is aimed at choosing a statistical predictive algorithm to fit in our analysis.

Among the considered techniques, the Random Forest and the Nearest Neighbor algorithms are resulted in the lowest error with high accuracy, in other words, if we take the area profile of the nearest-neighbor (the most common area profile of k-nearest-neighbors), that would give the right timeline type. The confusion matrix of the Random Forest classifier, and the precision, recall are estimated in the following Table 2. The receiver operating characteristic curve for visualizing the performance of the classifiers is described in Figure 9. This result shows that the area type *S5* is the well classified and compact by showing a strong correlation between the area activity categories and area timeline. The area types *S1*, *S2*, *S3* and *S4*, *S6* can be still refined in terms of the area activity categories.

Table 1: Results for the predictive models with the use of area types observed by spectral clustering algorithm

Algorithm	Cross Validation	Overall ACC
Random classifier	0.83	16.7%
Linear Discriminant	0.5404	45.01%
Quadratic Discriminant	0.4649	52.90%
Naive Bayes(kernel density)	0.6748	20.38%
K-NN (k=5, euclidean dist)	0.3822	61.73%
K-NN (k=10, euclidean dist)	0.4068	59.26%
Decision Tree	0.4806	52.58%
Random Forest	0.3513	64.89%
Multi-class SVM	0.4997	49.47%

Table 2: Confusion matrix and precision, recall and f-measure in each area type defined for predicting timeline based on location context about categorical human activity by Random Forest classifier

Area type defined	S1	S2	S3	S4	S5	S6	Prec.	Recall.	F-measure.
S1	8.91%	0.20%	1.80%	4.47%	0.07%	1.57%	52.35%	57.30%	54.71%
S2	0.10%	8.58%	0.70%	1.77%	0.47%	1.30%	66.41%	76.26%	70.99%
S3	1.77%	0.43%	10.15%	1.70%	1.27%	1.23%	61.29%	60.32%	60.80%
S4	2.34%	0.53%	1.13%	19.63%	0.33%	1.54%	76.96%	63.64%	69.67%
S5	0.03%	0.23%	1.37%	0.53%	6.54%	0.07%	74.52%	74.81%	74.67%
S6	2.40%	1.27%	1.67%	2.74%	0.07%	11.08%	57.64%	66.00%	61.54%

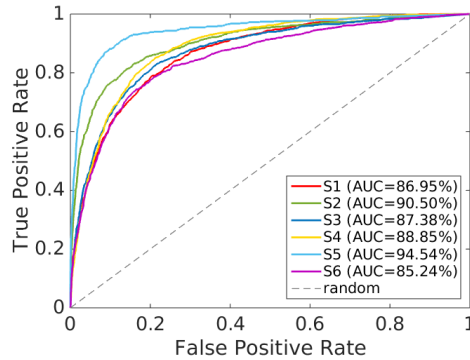


Figure 9: Receiver operating characteristic to multi-class by Random Forest classifier

5.2 Land-use type vs human behavior

After we observed strong prediction accuracy of timelines based on categorical area types, we analyze the relation between the timelines and land-use types which are for-

mally defined by land-use management organizations. While many works try to predict activity based on land-use, we perform a comparative study of the two approaches. We identify that even the area of the same land-use might have different area types in terms of area profiles and those are still different in terms of human activity timelines quantified through mobile phone records, which validates significance of activity-based classification vs official land-use. We predict the timeline type of a given area based on the land-use type using the Random Forest and the Nearest Neighbor classifiers. We used the land-use types from the OSM³ for this prediction task (see the distribution of land-use types of Milan in Figure 10(a)). The prediction accuracy of the Random Forest classifier is 53.47%. This shows that predicting power of categorical types is higher compared to land-use types.

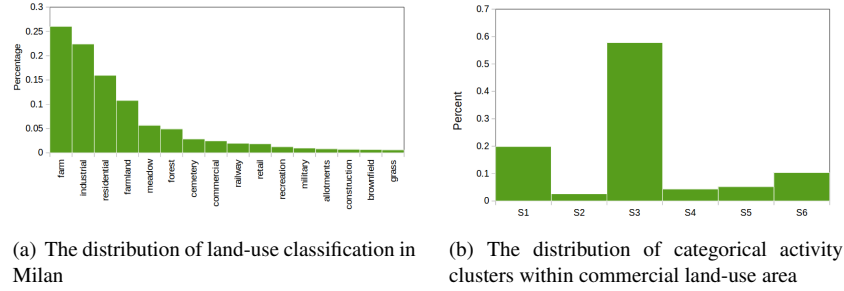


Figure 10

We also match the area types we observed with the land-use types given officially. The result shows that even within the same land-use type, the timelines corresponding to different clusters are still different. For example, 58% of the commercial land-uses matched with the area type S3 which followed by S1, S6 and S2, S3, S4, S5, shown in Figure 10(b). The corresponding timelines to the different clusters within the commercial land-use type are illustrated in Figure 11.

The timelines in the same area type observed, also in the same land-use officially defined, can be still refined, but the timeline pattern refinement will require more emphasis on the appropriate features, for example, timelines for weekday or weekend. The area profiles are semantically different concepts in terms of human activities performed in geographical areas. Further, it will allow us to identify a standard or exceptional type of mobile network activities in relevant areas, as well as to enable the identification of unknown correlations, or hidden patterns about anomalous behaviors.

³<http://wiki.openstreetmap.org/wiki/Key:landuse>

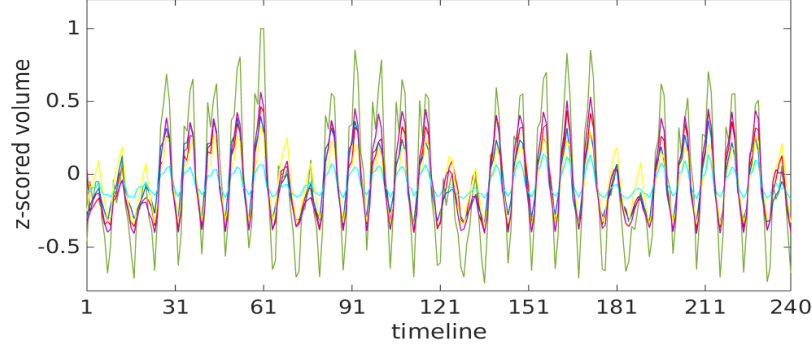


Figure 11: The timelines belong to different clusters within the commercial land-use: $S1$ is red, $S2$ is lime, $S3$ is blue, $S4$ is yellow, $S5$ is cyan/aqua, and $S6$ is magenta/fuchsia

6 Conclusion and Future works

In this paper, we proposed an approach that characterizes and classifies geographical areas based on their anticipated (through POI distribution) human activity categorical types, such as working or shopping oriented areas. We concentrated on the analysis of the relationship between such spatial context of the area and observed human activity. Our approach compares the similarity between area activity categorical profiles and human activity timeline categories estimated through cell phone data records. We found an overall correlation of 61% and canonical correlation of 65% between contextual and timeline-based classifications. We observed six types of areas according to the area activity categories where we compared their human activity timelines with their area activity categories and the correlation (canonical) coefficient is between 72% and 98%. For example, the area type $S5$ related to working activity has a strong correlation of 98% which followed by the area types, $S2$ related to sporting activity and $S3$ related to the human activities in the center of the city. The supervised learning approach validates possibility of using an area categorical profile in order to predict to some extent the network activity timeline (i.e., call, sms, and internet). For example, the Random Forest approach performs well with the accuracy of 64.89%. So human behaviors' temporal variation is characterized similarly in relevant areas, which are identified based on the categories of human activity performed in those locations. Furthermore we found that the prediction accuracy based on the official land-use types is only 53.47%. So the official land-use types by themselves are not enough to explain the observed impact of area context on human activity timelines, also because even within

the same land-use type, different activity categorical types still demonstrate different activity timelines. Further, the semantic description of area profiles associated to mobile phone data enables the investigation of interesting behavioral patterns, unknown correlations, and hidden behaviors in relevant areas. We expect the approach to be further applicable to other ubiquitous data sources, like geo-localized tweets, foursquare data, bank card transactions or the geo-temporal logs of any other service.

7 Acknowledgments

The authors would like to thank the Semantic Innovation Knowledge Lab - Telecom Italia for publicly sharing the mobile phone data records which were provided for Big Data Challenge organized in 2013, Italy. We also would like to thank MIT SENSEable City Lab Consortium partially for supporting the research.

References

- [1] Canonical correlation analysis. In *Applied Multivariate Statistical Analysis*, pages 321–330. Springer Berlin Heidelberg, 2007.
- [2] J. P. Bagrow, D. Wang, and A.-L. Barabási. Collective response of human populations to large-scale emergencies. *CoRR*, abs/1106.0560, 2011.
- [3] R. A. Becker, R. Cceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. A tale of one city: Using cellular network data for urban planning.
- [4] F. Calabrese, P. F. C., G. Di Lorenzo, L. Liu, and C. Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *the Proc. of the 8th international conference on Pervasive Computing*, (Pervasive’10), pages 22–37, Berlin, Heidelberg, 2010. Springer-Verlag.
- [5] F. Calabrese and C. Ratti. Real time rome. *Networks and Communication studies*, 20(3-4):247–258, 2006.
- [6] Z. Dashdorj and L. Serafini. Semantic enrichment of mobile phone data records using linked open data. In *the Proc. of the 12th Intl. Conf. Semantic Web Conference Poster and Demonstrations Track*, 2013.
- [7] Z. Dashdorj and L. Serafini. Semantic interpretation of mobile phone records exploiting background knowledge. In *the Proc. of the 12th Intl. Conf. Semantic Web Conference Doctoral Consortium*, 2013.

- [8] Z. Dashdorj, L. Serafini, F. Antonelli, and R. Larcher. Semantic enrichment of mobile phone data records. In *MUM*, page 35, 2013.
- [9] Z. Dashdorj and S. Sobolevsky. Impact of the spatial context on human communication activity. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, UbiComp '15, pages 1615–1622, New York, NY, USA, 2015. ACM.
- [10] Z. Dashdorj, S. Sobolevsky, L. Serafini, F. Antonelli, and C. Ratti. Semantic enrichment of mobile phone data records using background knowledge. *arXiv preprint arXiv:1504.05895*, 2015.
- [11] Z. Dashdorj, S. Sobolevsky, L. Serafini, and C. Ratti. Human activity recognition from spatial data sources. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*, MobiGIS '14, pages 18–25, New York, NY, USA, 2014. ACM.
- [12] K. Farrahi and D. Gatica-Perez. What did you do today? discovering daily routines from large-scale mobile data. In *ACM International Conference on Multimedia (ACMMM)*, 0 2008. IDIAP-RR 08-49.
- [13] V. Frías-Martínez, V. Soto, H. Hohwald, and E. Frías-Martínez. Characterizing urban landscapes using geolocated tweets. In *SocialCom/PASSAT*, pages 239–248, 2012.
- [14] T. Fujisaka, R. Lee, and K. Sumiya. Exploring urban characteristics using movement history of mass mobile microbloggers. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, HotMobile '10, pages 13–18, New York, NY, USA, 2010. ACM.
- [15] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Identifying users profiles from mobile calls habits. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, UrbComp '12, pages 17–24, New York, NY, USA, 2012. ACM.
- [16] F. Girardin, F. Calabrese, F. D. Fiore, C. Ratti, and J. Blat. Digital footprinting: Uncovering tourists with user-generated content. *Pervasive Computing, IEEE*, 7(4):36–43, 2008.
- [17] F. Girardin, A. Vaccari, R. Gerber, and A. Biderman. Quantifying urban attractiveness from the distribution and density of digital footprints. *Journal of Spatial Data Infrastructure Research*.

- [18] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti. Towards a comparative science of cities: using mobile traffic records in New york, London and Hong Kong. *CoRR*, abs/1406.4400, 2014.
- [19] J. Han, M. Kamber, and A. K. H. Tung. *Spatial Clustering Methods in Data Mining: A Survey*. Taylor and Francis, 2001.
- [20] B. Hawelka, I. Sitko, E. Beinart, S. Sobolevsky, P. Kazakopoulos, and C. Ratti. Geo-located twitter as proxy for global mobility pattern. *Cartography and Geographic Information Science*, pages 1–12, 2014.
- [21] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.
- [22] C. Kang, S. Sobolevsky, Y. Liu, and C. Ratti. Exploring human movements in singapore: a comparative analysis based on mobile phone and taxicab usages. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 1. ACM, 2013.
- [23] J. Ni and C. V. Ravishankar. Pointwise-dense region queries in spatio-temporal databases. In R. Chirkova, A. Dogac, M. T. zsu, and T. K. Sellis, editors, *ICDE*, pages 1066–1075. IEEE, 2007.
- [24] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *The Social Mobile Web*, 2011.
- [25] S. Paldino, I. Bojic, S. Sobolevsky, C. Ratti, and M. C. González. Urban magnetism through the lens of geo-tagged photography. *arXiv preprint arXiv:1503.05502*, 2015.
- [26] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28(9):1988–2007, 2014.
- [27] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, and C. Zhou. A new insight into land use classification based on aggregated mobile phone data. *CoRR*, abs/1310.6129, 2013.
- [28] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. Activity-aware map: identifying human daily activity pattern using mobile phone

- data. In *the Proc. of the 1st Intl. Conf. Human Behavior Understanding*, pages 14–25, 2010.
- [29] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone location data. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 971–976, 2010.
 - [30] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz. Redrawing the map of great britain from a network of human interactions. *PLoS One*, 5(12):e14248, 2010.
 - [31] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli. Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and planning B*, 33(5):727, 2006.
 - [32] C. Ratti, S. Williams, D. Frenchman, and R. M. Pulselli. Mobile Landscapes: using location data from cell phones for urban analysis. *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN*, 33(5):727, 2006.
 - [33] J. Reades, F. Calabrese, and C. Ratti. Eigenplaces: analysing cities using the space time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009.
 - [34] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular Census: Explorations in Urban Data Collection. *IEEE Pervasive Computing*, 6(3):30–38, 2007.
 - [35] G. Sagl, E. Beinat, B. Resch, and T. Blaschke. Integrated geo-sensing: A case study on the relationships between weather and mobile phone usage in northern italy. In *IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services, ICSDM 2011, Fuzhou, China, June 29 - July 1, 2011*, pages 208–213, 2011.
 - [36] G. Sagl, T. Blaschke, E. Beinat, and B. Resch. Ubiquitous geo-sensing for context-aware analysis: Exploring relationships between environmental and human dynamics. *Sensors*, 12(7):9800–9822, 2012.
 - [37] P. Santi, G. Resta, M. Szell, S. Sobolevsky, S. Strogatz, and C. Ratti. Taxi pooling in New York City: a network-based approach to social sharing problems. *arXiv preprint arXiv:1310.2963*, 2013.
 - [38] S. Sobolevsky, I. Sitko, R. T. D. Combes, B. Hawelka, J. M. Arias, and C. Ratti. Money on the move: Big data of bank card transactions as the new proxy for

- human mobility patterns and regional delineation. the case of residents and foreign visitors in spain. In *Big Data (BigData Congress), 2014 IEEE International Congress on*, pages 136–143. IEEE, 2014.
- [39] S. Sobolevsky, I. Sitko, S. Grauwin, R. T. d. Combes, B. Hawelka, J. M. Arias, and C. Ratti. Mining urban performance: Scale-independent classification of cities based on individual economic transactions. *arXiv preprint arXiv:1405.4301*, 2014.
 - [40] S. Sobolevsky, M. Szell, R. Campari, T. Couronn, Z. Smoreda, and C. Ratti. Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLoS ONE*, 8(12):e81707, 12 2013.
 - [41] V. Soto and E. Frías-Martínez. Robust land use characterization of urban landscapes using cell phone data, June 2011. The First Workshop on Pervasive Urban Applications (PURBA).
 - [42] V. Soto and E. Frías-Martínez. Robust land use characterization of urban landscapes using cell phone data, 2011.
 - [43] T. Stathopoulos, H. Wu, and J. Zacharias. Outdoor human comfort in an urban climate. *Building and Environment*, 39(3):297 – 305, 2004.
 - [44] J. Sun, J. Yuan, Y. Wang, H. Si, and X. Shan. Exploring spacetime structure of human mobility in urban space. *Physica A: Statistical Mechanics and its Applications*, 390(5):929–942, 2011.
 - [45] P. Tucker and J. Gilliland. The effect of season and weather on physical activity: A systematic review. *Public Health*, 121(12):909 – 922, 2007.
 - [46] M. R. Vieira, V. Frias-Martinez, N. Oliver, and E. Frias-Martinez. Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM '10*, pages 241–248, Washington, DC, USA, 2010. IEEE Computer Society.
 - [47] U. von Luxburg. A tutorial on spectral clustering. *CoRR*, abs/0711.0189, 2007.
 - [48] S. Wakamiya, R. Lee, and K. Sumiya. Urban area characterization based on semantics of crowd activities in twitter. In *Proceedings of the 4th international conference on GeoSpatial semantics, GeoS'11*, pages 108–123, Berlin, Heidelberg, 2011. Springer-Verlag.

- [49] Q. Wang and J. E. Taylor. Quantifying human mobility perturbation and resilience in hurricane sandy. *PLoS ONE*, 9(11):e112608, 11 2014.
- [50] Y. Yuan and M. Raubal. Extracting dynamic urban mobility patterns from mobile phone data. In N. Xiao, M.-P. Kwan, M. F. Goodchild, and S. Shekhar, editors, *GIScience*, volume 7478 of *Lecture Notes in Computer Science*, pages 354–367. Springer, 2012.